Empirical Methods in Artificial Intelligence: A Review

Pat Langley

Early research on AI typically involved qualitative demonstrations of intelligent behavior, with novelty being the primary focus. However, as the field has matured, there have been increasing demands for more careful evaluation using quantitative measures of behavior. In some cases, the response has taken the guise of formal analyses, and in others, it has emphasized comparisons between system and human behavior, but the predominant movement has been toward empirical studies of AI methods. As a result, techniques for experimental design, exploratory data analysis, and statistical testing, originally developed in other fields, have become increasingly relevant for AI researchers.

Paul Cohen's book *Empirical Methods for Artificial Intelligence* aims to encourage this trend by providing AI practitioners with the knowledge and tools needed for careful empirical evaluation. The volume provides broad coverage of experimental design and statistics, ranging from a gentle introduction of basic ideas to a detailed presentation of advanced techniques, often combined with illustrative examples of their application to the empirical study of AI. The book is generally well written, clearly organized, and easy to understand; it contains some mathematics—but not enough to overwhelm readers. Examples come from AI work on planning, machine learning, natural language, and diagnosis.

The text includes introductory chapters on the nature of empirical research, methods for exploratory data analysis, and issues in experimental design. The ensuing chapters cover analytic and computer-intensive techniques for statistical inference, including schemes for hypothesis testing and parameter estimation. Next come treatments of performance assessment (that is, the measurement of system behavior) and analysis of variance, including methods for detecting interactions among variables. A chapter on modeling covers the use of linear regression and related procedures in characterizing system behavior, and the final chapter presents tactics for generalizing from empirical results.

One of Cohen's most important points concerns the goals of empirical AI research. Repeatedly, he asks what question an experimental study or test was designed to answer. His generic stance is that such studies should help one understand the mapping from characteristics of AI methods, domains, and tasks onto characteristics of system behavior. He introduces this idea at the outset and then uses examples to drive the point home, continually reminding readers of its import. The goal of empirical research is not to determine the winner in a competition but to gain understanding of the reasons for differences in behavior. This view is typically lacking even in highly empirical subfields such as machine learning, and readers would do well to follow Cohen's lead on this front.

In my view, the book's conceptual contributions are more important than its coverage of statistical techniques. One can find a recipe for analysis of variance or correlation in nearly any statistics text, but it is much more difficult to find a clear introduction to basic concepts of experimental method, such as control conditions, ceiling effects, sampling bias, and order effects. Cohen's coherent presentation of these subtle ideas, and his illustration of them in AI contexts, makes them seem obvious. However, these terms almost never appear in the experimental AI literature, suggesting that few authors are aware of their importance. The book also includes useful heuristics for experimental design, such as the hints for designing factorial experiments in Chapter 3 and the techniques for determining sample size in Chapter 4. In addition, the closing chapter contains insightful advice about generalizing beyond one's empirical results.

Despite its strengths, the book does have some drawbacks. The text typically presents an abstract method and then illustrates it with an AI example, but the examples are not always present. Moreover, Cohen draws too many examples from his own work on planning and a few other favorites, giving the book less AI coverage than it might have. The volume also gives too much attention to statistical hypothesis testing, which emphasizes a particular type of question (whether two behaviors are different) that is not always the most illuminating. Fuller treatment of methods for understanding the source of an AI system's power, such as lesion studies and parametric experiments, would have served better.

Even the book's breadth has disadvantages, in that it covers many statistical techniques, such as multiple linear regression, that are unlikely to see wide use within AI in the near future. As a result, few instructors will select the book as a text, and few researchers will read the entire volume. Given the state of the field's knowledge about experimental design and statistics, a more realistic volume would have focused on more basic methods, such as those covered in the first few chapters, and dealt with a broader set of AI examples to ensure relevance to readers' interests.

Nevertheless, *Empirical Methods in Artificial Intelligence* is an excellent reference text that covers a wide range of issues in a careful and thoughtful manner. Many researchers will want it on their shelves for use when they decide some new design or test is needed to evaluate their work, and they will want to encourage students to read substantial portions of the volume. In many ways, this book is ahead of its time, but its presence should help accelerate the evolution of AI toward a rigorous yet relevant empirical science.

Pat Langley holds a research position at Stanford University and serves as director of the Institute for the Study of Learning and Expertise.